CS 124 Machine Translation Report

Aditya Sarkar – sarkar17@stanford.edu
Raymond Chan – rchan2@stanford.edu
Travis Le – tle16@stanford.edu
Tanner Gilligan – tanner12@stanford.edu

February 28, 2015

1 Introduction

For our machine translation assignment, we chose to translate French to English. While none of us are fluent in the language, two of us did have some experience from high-school. French was challenging for us to translate for several reasons, most notably the structural differences between the languages, and the fact that many direct translations simply dont make sense in English. One structural difference we noted is Frenchs noun-adjective ordering (vs. Englishs adjective-noun ordering). An example of this is the phrase ”mre biologique”, which directly translates to ”mother biological”, as opposed to the correct translation of ”biological mother”. Another structural difference is Frenchs use of reflexives, which often has a reversed object-verb ordering (as opposed to Englishs verb-object ordering). An example of this is the phrase ”me bloquait”, which directly translate to ”me block”, as opposed to the correct translation of ”block me”. In addition, there are some phrases in French that, when translated to English, simply dont make sense. Other issues we encountered include the fact that nouns in French have a gender, and that a single French word, such as the pronouns ”il” and ”elle”, can be translated to many different English words. In addition to these lingual problems, we also had technical problems due to French. Since French uses accented letters, we spent quite a bit of time trying to get both our code and our external libraries to work with these non-Ascii characters. Below are our dev and test sets:

2 Development Set

- Il est poursuivi pour assassinats et risque la peine de mort.
- Ils encourent une interdiction vie.
- Le football doit lutter au quotidien contre le racisme.
- L’entreprise augmente le salaire d’un demi-million d’employs.
- Il rejoint Bordeaux, o il retrouve sa mre biologique, frquente le monde de la nuit et se rapproche de la pgre locale.
- Une tranche de cette monstruosit contient elle seule 450 calories.
- Mais la surprise des mdecins, elle a dvelopp une obsit.
- Ce petit drone au petit physique est dot de trois helices et d’un appareil photo.
- J’ai voulu entrer dans le wagon mais un groupe de supporteurs anglais me bloquait et me repoussait.
- Les lus, de gauche comme de droite, sont pourtant favorables cette concurrence.
- Il se librera de son corps, ne fera plus qu’un avec l’ordinateur et, grce l’intelligence artificielle, accdera l’immortalit.
- Longtemps, la majeure partie de la communaut scientifique est reste sans raction face aux thses des transhumanistes, qu’elle jugeait peu crdibles.

3 Test Set

- Ces graisses satures se trouvent principalement dans les viandes grasses.
- La pizza est galement surmonte de bacon et de pepperoni.
- Il y a des dizaines de fouilles prventives en France chaque anne.
- La pollution s'attaque leurs chances de bien se reproduire.
- Le transhumanisme existait avant l'explosion des hautes technologies.

4 Pre-Processing

One of the pre-processing strategies that we implemented was to add part of speech tags for the french words, and to reorder the adjectives and nouns of the sentences. We implemented this to address the fact that French has noun-adjective ordering, while English has adjective-noun ordering. We can find the following examples of this noun-adjective ordering in the dev set: "fats saturated", "mother biological", "supporters English", "intelligence artificial", and "community scientific". After implementing this strategy, we were able to reverse this ordering of these phrases, resulting in more intelligible English.

Due to the fact that many French words translate to multiple English words, one of our problems was making our translator select the correct translation, especially for pronouns. When we originally constructed our dictionary, we simply wrote down the translations in the order Google Translate provided them based on their frequency. The problem with this is that Google only provides 3 different frequency levels, so its difficult to decide the ordering within each level. In order to rectify this, we looked through transcripts of the European Parliament which contains fluent translations in French and English. Using these transcripts, we were able to reorder the translation list to more accurately reflect the true frequencies of words. For example, when we translated "elle" using Google Translate, we found that "it" and "she" were 2 equally likely translations, so we put them into our dictionary in that order. However, after looking through some of the transcripts, we realized that "elle" nearly always refers to "she", and very rarely refers to "it". In order to account for this, we flipped "it" and "she" in our translation list.

Another pre-processing strategy we made use of was flipping the subject and verb in reflexive verbs. Reflexive verbs, otherwise known as prenominal verbs, are used in French roughly when the subject performing the action (verb) is the same as the object being acted upon. This is not really something that happens in English, and hence leads to problems with the ordering of the words. For example, the French phrase "Il me bloque" correctly translates to "he blocked me" in English, however, word for word, it translates to "he me block". While it sounds humorous, it is obviously incorrect. To fix this, we use our French word-tags to detect when a reflexive pronoun is followed by a verb, and then switch them in order to generate intelligible English.

A pre-processing strategy we made use of was expanding French contractions. With certain words (usually pronouns), if the last letter of the previous word and the first letter of the next word are both vowels, they contract the two words using an apostrophe. For example the phrase "I am called", "Je me appelle" is actually spelled "Je mappelle". The reason we expanded these contractions is so that when we put words in the dictionary, they were more general and hence could be used with multiple conjugations. In addition, this allows the POS tagger to correctly identify every part of the sentence, and also allows us to use the expanded words in our other processing strategies. We also found that Google translate does not deal well with contracted words, and would therefore alter our dictionary. An example of this is the French phrase "de un" versus its correct spelling of "dun". In both cases, the correct translation is "of a", however Google Translate translates "de un" to "of a", but translates "dun" to "a".

5 Post-Processing

Another problem we encountered in regards to reflexive verbs is the use of the word "se". For example, the French phrase "ils se douchent" translates to (with the word ordering corrected for the purpose of this explanation) "they shower themselves". In English, we see the "themselves", which is a literal translation of "se", as superfluous. Therefore, in order to produce more fluent English, we removed the word "se" from French sentences as a pre-processing step. The only exception to this removal is the case where "se" is preceded by "il" or "elle" since these cases produce a different use of "se".

We actually further post processed on these excpeted ”se”s. ”Se” preceded by ”il” means himself, while ”se” preceded by ”elle” means herself. This translation is usually needed in the English version of the sentence, while ”themselves” should nearly always be omitted to produce intelligible English.

Another strategy we implemented was based on how French and English treat plurals in the present tense. Lets consider the two phrases: ”The men eat” and ”the man eats”. Here, when the noun is singular, the verb is singular and vice versa. In french this is dealt with more normally, plural nouns are followed by plural versions of verb conjugations (have an s or similar at the end). This gives us a problem when we directly translate, so we used our POS tagger to find nouns followed by verbs in our translations and we pluralised any present tense participle of a verb if it was preceded by a singular noun or pronoun.

A large difference we noted in our dev set is the fact that French makes liberal use of articles, namely that they include articles where there are none in English. An example of this is the French phrase ”lutter au quotidien contre le racisme”, which directly translates to ”fight the daily against the racism”. A post-processing strategy we made use of to correct this was to compare trigrams/bigrams by querying Googles Ngram Viewer. Our strategy was as follows: in the case that in some part of the sentence we come across a fragment like ”X (to—the—a) Y”, we want to see if (to—the—a) should be there. We do this by comparing the trigram frequency of the original fragment, ”X (to—the—a) Y”, with the bigram frequency of ”X Y”, and select the bi/tri-gram with the higher probability. We reason that, if the bigram frequency of ”X Y” is greater than the trigram frequency of ”X (to—the—a) Y”, then the middle word may be a weird grammatical translation mistake. There are two exceptions to this, however. If the middle word is ”the” and the X word is (”of” or ”to), then we do not do this comparison for removal. This is because ”to the” and ”of the” are very common phrases in English.

6 Google Translate Comparison

1. French: Ces graisses satures se trouvent principalement dans les viandes grasses.
   Google: These saturated fats are found mainly in fatty meats.
   Ours: These saturated fats find mainly in fat meat.
   Comparison: Googles translations is clearly superior to ours, specifically since it was able to correctly identify the tense of the verb find, and was able to identify the fact that ”meat” is plural.

2. French: La pizza est galement surmonte de bacon et de pepperoni.
   Google: The pizza is topped with bacon and pepperoni.
   Ours: The pizza is also overcome of bacon and of pepperoni.
   Comparison: Googles translations is once again superior, due to its ability to parse ”est galement surmonte de” as ”is topped with”. This difference is largely due to the fact that our translation program directly translates the sentence, then reorders and drops words. This doesnt give us the ability to combine 2 French words into a single english word, as would be required for a correct translation.

3. French: Il y a des dizaines de fouilles prventives en France chaque anne.
   Google: There are dozens of rescue excavations in France each year.
   Ours: He there of decades of preventative excavation in France each year.
   Comparison: Googles translation is clearly more fluent than ours, and produced a more correct translation. This is again due to the fact that we have to facilities for combining 2 French words into 1 English word, which would be required to translate ”Il y a des” to ”There are”.

4. French: La pollution s’attaque leurs chances de bien se reproduire.
   Google: Pollution attacks their chances of successful reproduction.
   Ours: The pollution attacks their chances of good reproduce.
   Comparison: Googles translation is close to ours, but was able to better select synonyms (”good” vs ”successful”) as well as omit the initial ”La” from the sentence.

5. French: Le transhumanisme existait avant l’explosion des hautes technologies.
   Google: Transhumanism existed before the explosion of high technologies.
   Ours: The transhumanism exists before the explosion of high technologies.
   Comparison: For once, our translation is nearly identical to Googles, the only difference being that Google was able to omit the initial ”Le” from the sentence, which we didnt account for since it only occurred once in our dev set.

7 Error Analysis

7.1 Error 1

One of the recurrent errors was the inaccuracy of the tense in our results. Sometimes the sentence accurately reflect the correct tense, but usually it wouldnt. This is because our dictionary was built using a word-by-word translation through Google Translate. The problem with Google Translate though is that, without context around a verb, it will almost always translate it to the present tense regardless of the original tense of the verb.
Examples: "these saturated fats find mainly in fat meat" (Should be found)
"the transhumanism exists before the explosion of high technologies" (Should be existed)
Solution: In order to solve this issue, we could construct a table of all conjugations for each French verb, and map it to its corresponding English translation based on the context of the French sentence. Even though its not always a 1-to-1 mapping, this would lead to improved outputs.

7.2 Error 2

Another error that was common was the incorrect use of articles, prepositions, and pronouns, as well as our inability to combine them. This happened due to the fact that we started off with a direct 1-to-1 translation, and only looked for excessive uses of the words "to", "the", and "a".
Examples: Even though we tried to omit excessive articles, we were still not 100% successful as is apparent in the example: "The transhumanism exists before the ".
An example of our inability to combine them is the phrase: "He there of decades of preventative".
A lot of the times in French, multiple words are used where English would only use one. For example "Il y" has a correct translation of "there" but there is no way for us to condense two french words into one english word.
Solution: Our solution for this would be to try and create some sort of 2-to-1 mapping for French to English. This would allow us to combine French prepositions/pronouns/articles into intelligible English.

7.3 Error 3

French nouns are always preceded with an article. In english, the first word of the sentence could be a noun but in French the noun must be preceded by an article unless it is proper. The reason for this is similar, to the reason for error 2. When we translate word for word, we have to directly translate the preceding article first and we end up with it in the English translation.
Examples: "The pollution attacks their chances of good reproduce" (Should be found)
"The transhumanism exists before the explosion of high technologies" (Should be existed)
Solution: We would need a better mapping structure than our current one-to-one mapping. We could use bigram probabilities to figure out if the article is actually needed before the noun, which we did try this to a certain extent, but there is still plenty of room for improvement.

7.4 Error 4

An error that occurs once in the test set, but would likely occur frequently in a larger test set, is the case of adjectives preceding verbs. This probably happens because, in French, when an adjective is followed by a noun, the form of the noun is actually also a form of the verb (reproduce vs reproduction). Google translates this noun to the english infinitive verb form though.
Examples: "... attacks their chances of good reproduce."
Solution: We should have implemented post processing check where we changed any verb following and adjective to the noun form of that verb.